Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Explainable Machine Learning for Predicting Lactococcus Abundance in Rump Steak Enhanced by Generative AI-Based Metagenomic Data

Volkan Ince
University of Portsmouth
Mohamed Bader-El-Den
Mohamed.Bader@port.ac.uk
University of Portsmouth
Omer Faruk Sari
University of Portsmouth
Annette Sansom
Campden BRI (United Kingdom)

Research Article

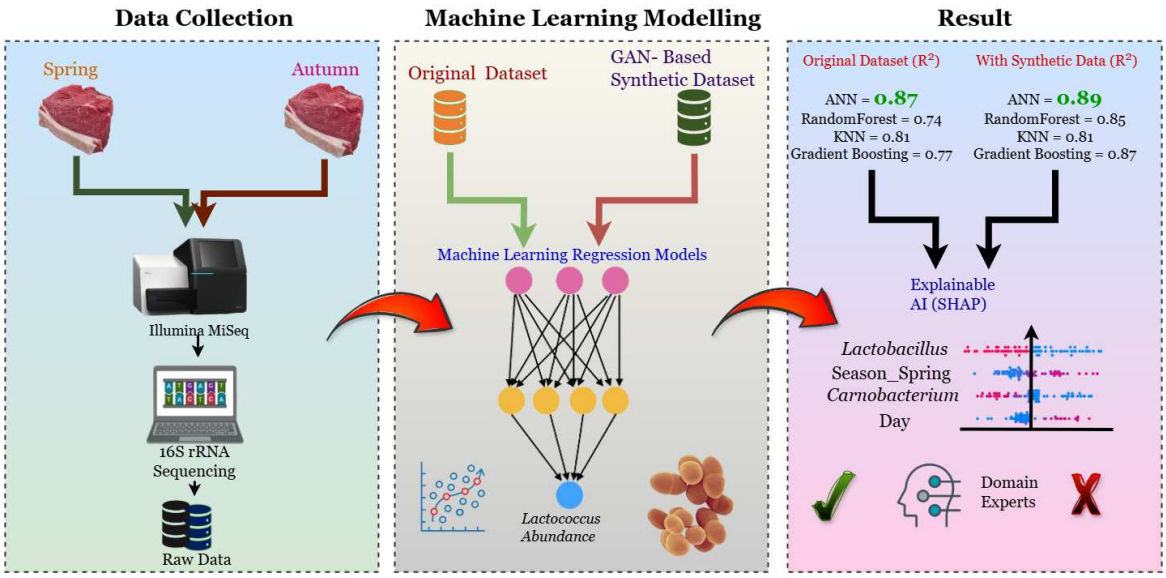
Keywords: Machine learning, Generative AI, Explainable AI, Metagenomic data, Lactococcus

Posted Date: April 28th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-6515642/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.



Graphical Abstract

Explainable Machine Learning for Predicting Lactococcus Abundance in Rump Steak Enhanced by Generative AI-Based Metagenomic Data

Volkan Ince[1†], Mohamed Bader-El-Den[1*†], Omer Faruk Sari[1], Annette Sansom[2]

[1]Computer Science, University of Portsmouth, Portsmouth, PO1 3GY, UK.
[2]Emerging Microbiology, Campden BRI, Gloucestershire, GL55 6LD, UK.

*Corresponding author(s). E-mail(s): Mohamed.Bader@port.ac.uk;
Contributing authors: volkan.ince@port.ac.uk; omer.sari@port.ac.uk;
Annette.Sansom@campdenbri.co.uk;
†These authors contributed equally to this work.

Abstract

The microbial spoilage of meat products poses a significant challenge to food safety and shelf-life, primarily due to the proliferation of spoilage-related bacteria under varying environmental conditions. Among beneficial microorganisms, *Lactococcus lactis* species are notable for their ability to produce nisin, a natural antimicrobial that enhances the microbial stability of meat. This study presents an interpretable machine learning framework to predict the relative abundance of *Lactococcus* in meat using 16S rRNA sequencing data, enhanced by synthetic data generation and explainable AI. A dataset comprising microbial profiles from rump steak samples across seasonal and temporal variables was expanded using a Generative Adversarial Network (GAN), addressing data scarcity. Multiple regression models, including Artificial Neural Networks, Random Forest, Gradient Boosting, and k-Nearest Neighbors, were trained and evaluated. The ANN model achieved the highest performance ($\boldsymbol{R^2 = 0.89}$) when trained on synthetic data and tested on real samples. Explainable AI analysis revealed *Lactobacillus*, *Carnobacterium*, and seasonal variables as the most influential predictors. Notably, seasonal shifts significantly impacted *Lactococcus* abundance, with higher levels observed in autumn. This study demonstrates the effectiveness of GAN-augmented data and interpretable machine learning in modeling microbial behavior, providing a scalable, cost-efficient tool for predictive microbiology. The approach not only enhances our understanding of microbial co-occurrence patterns but also supports the development of intelligent, data-driven strategies for natural food preservation.

Keywords: Machine learning, Generative AI, Explainable AI, Metagenomic data, Lactococcus

1 Introduction

Meat products are highly perishable food items, and their quality deterioration during storage is primarily attributed to microbial proliferation and metabolic activity (Nath et al. 2023; Kim et al. 2024; Zhang et al. 2024). The microbial composition of meat products significantly influences their quality, shelf life, and safety (Toomik et al. 2023; Qu et al. 2024). Among the spoilage-causing microorganisms, *Pseudomonas spp* (Li et al. 2014; Spanu et al. 2018)., *Brochothrix thermosphacta* (Mohsina et al. 2020), *Carnobacterium spp.* (Wang et al. 2023), and certain lactic acid bacteria (LAB) are predominant contributors to the deterioration of meat products. However, some LAB strains exhibit antagonistic activity against spoilage and pathogenic bacteria by producing organic acids,

hydrogen peroxide, and bacteriocins, thereby enhancing the microbial stability and safety of meat products (Marcelli et al. 2024). Among these beneficial microbes, *Lactococcus* species are particularly noteworthy for their ability to synthesize nisin, a well-characterized bacteriocin with strong inhibitory effects against Gram-positive spoilage and pathogenic bacteria (Zhao et al. 2016; Chaves López et al. 2017; Bi et al. 2020; Kuda et al. 2024). The production of nisin by *Lactococcus* not only extends the shelf life of meat products but also contributes to their microbiological safety, positioning it as a promising agent in natural food preservation strategies (Phuapaiboon et al. 2013; Casco et al. 2022).

The identification and characterization of microbial communities in meat products have been greatly advanced by high-throughput sequencing technologies. In particular, 16S rRNA gene metabarcoding has enabled detailed taxonomic profiling of complex microbial ecosystems (Watanabe et al. 2023; Piredda et al. 2022; Merchán et al. 2024). While these technologies offer a wealth of compositional data, interpreting the functional and contextual implications of microbial shifts remains a major challenge. Raw sequencing data alone cannot reveal causality or predict microbial dynamics under varying production or environmental conditions.

To overcome this limitation, recent research has increasingly focused on predictive microbiology (Koyama et al. 2022; Yildirim-Yalcin et al. 2025; Tarlak 2024; Oliveira et al. 2024; Al et al. 2024), which models microbial behavior as a function of intrinsic (e.g., pH, water activity) and extrinsic (e.g., temperature, storage time) factors. In this context, machine learning (ML) and artificial intelligence techniques have gained prominence for their capacity to capture complex microbial dynamics under diverse environmental conditions. These approaches offer significantly improved predictive accuracy compared to traditional parametric models, particularly in forecasting microbial growth, inactivation, and interspecies competition (Yücel and Tarlak 2023; Koyama et al. 2022; Tarlak 2024). However, the complexity of ML models often limits transparency, which can be a concern in sensitive domains like food safety. To that end, explainable artificial intelligence methods—particularly SHapley Additive exPlanations (SHAP)—have been developed to enhance model transparency and facilitate meaningful interpretation of complex outputs (Lundberg and Lee 2017).

Despite the promise of these approaches, a recurring challenge in applying ML to microbiome datasets is the issue of limited sample size (Papoutsoglou et al. 2023). Sequencing experiments are often constrained by financial, logistical, or time-related limitations, leading to relatively small datasets that are insufficient for training complex models (Ince et al. 2025). Small sample sizes increase the risk of overfitting, reduce model generalizability, and limit the statistical power of downstream analyses (Topçuoğlu et al. 2020). To deal with this constraint, recent studies have explored the use of generative models, such as Generative Adversarial Networks (GANs), to augment datasets with synthetic data (Arora and Arora 2022; Lu et al. 2022; Mert 2023).

While machine learning has become increasingly common in food microbiology, its application to modeling beneficial microbes in meat systems—particularly nisin-producing *Lactococcus lactis*—remains underexplored. Despite growing interest in ML-based predictive models, no studies to date have directly investigated the behavior or dynamics of L. lactis. Furthermore, due to the high cost and labor-intensive nature of microbial sampling and testing, the potential of synthetic data generation techniques, such as GAN-based augmentation, has been largely overlooked in this context. Equally important, the lack of explainable AI (XAI) applications represents a key gap in the literature, as these tools could significantly improve expert understanding and trust by making model predictions more transparent and actionable.

To address this gap, the present study combines 16S rRNA gene metabarcoding with supervised ML models and SHAP analysis to predict and interpret variations in *Lactococcus* abundance within a meat product. First, the relative abundance of *Lactococcus* was quantified using sequencing data. Then, we employed multiple machine learning algorithms—including Artificial Neural Networks (ANN), Random Forest, Gradient Boosting, and k-Nearest Neighbors (kNN) regression—to model its abundance based on seasonal variables (e.g., spring vs. autumn), production parameters (e.g., storage day), and the relative abundance of other microbial taxa. To overcome the challenge of limited sample size, a GAN was used to create a synthetic dataset that enriched the training data and improved model robustness. Finally, SHAP analysis was applied to interpret the model's output and to identify the most influential factors driving *Lactococcus* levels. By combining microbial sequencing data with predictive and explainable modeling techniques, this study contributes to the design of more informed, effective strategies for enhancing food safety and microbial quality.

Specifically, this study contributes to the field in the following ways:

- An interpretable model is built using 16S rRNA data, machine learning, and SHAP to predict *Lactococcus* abundance.
- Seasonal shifts, storage time, and microbial co-occurrence emerge as key drivers of *Lactococcus* dynamics.
- Explainable AI enables transparent insights from complex models in food microbiology.
- GAN-based synthetic data improves predictive accuracy under limited data conditions.

The paper is structured as follows: In Section 2, related work is reviewed to provide context and highlight existing gaps in the literature. In Section 3, we provide a detailed explanation of our proposed method. Following this, in Section 4, we present the experimental results and analyse the performance of the proposed methods. Finally, in Sections 5 and 6, we conclude the paper by summarizing our findings and suggesting new directions for future research.

2 Related Work

Recent advancements in ML have revolutionized predictive microbiology by offering powerful tools for modeling microbial dynamics under complex, multidimensional conditions. Unlike traditional kinetic models, ML algorithms can capture non-linear relationships and high-order interactions between environmental variables and microbial responses without relying on predefined mathematical structures. This capacity makes ML particularly suitable for high-throughput applications and large-scale risk assessments in food safety.

A number of studies have demonstrated the efficacy of ML regression models for predicting microbial growth under varying environmental conditions and food matrices. Koyama et al. (2022) applied SVR, ETR, and GPR to model *E. coli* O157:H7 growth without using primary or secondary models, with SVR achieving the highest performance ($R^2 = 0.93$, RMSE = 0.36). However, their model's interpretability was limited and dataset size was relatively small, affecting generalizability. Al et al. (2024) compared RF, ANN, SVR, and MLR for modeling *E. coli* O157:H7 growth in beef, where RF performed best ($R^2 = 0.98$, MSE = 0.08), but again, the study was constrained by a limited dataset. In contrast, Yücel and Tarlak (2023) used a more extensive dataset (n = 2654) from ComBase and demonstrated strong performance of RFR in modeling multiple pathogens in beef, with $R^2 = 0.949$, without explicitly reporting major limitations.

Complementary applications by Tarlak (2024) used machine learning models to model *Pseudomonas* spp. in culture media, identifying GPR as most accurate ($R^2_{\text{adj}} = 0.923$, RMSE = 0.005), though interpretability and dataset size were acknowledged as constraints. Tarlak and Yücel (2023) extended these approaches to multiple food matrices, showing RFR superiority ($R^2 = 0.913$), but similarly noted issues of transparency and dataset size. Yildirim-Yalcin et al. (2025) predicted total mesophilic bacteria in spinach using SVR, DTR, and GPR, where GPR performed best ($R^2 = 0.9888$, RMSE = 0.0811); yet, the model was trained on a relatively small dataset and lacked interpretability, limiting its robustness.

Growth/no-growth (G/NG) boundary modeling has also gained traction in food safety applications. Oliveira et al. (2024) developed a logistic regression model to predict *Clostridium sporogenes* behavior in processed cheese using pH, NaCl, and nisin as predictors. Despite identifying key inhibitory conditions, their model accuracy was moderate (62.8%) and limited by a small dataset (n = 75). Sun et al. (2021) achieved high predictive accuracy ($c - \text{values} > 0.98$) for *Bacillus* spores in acidic environments using 1920 binary observations, with no critical limitations reported. Meanwhile, Kuroda et al. (2019) employed deep learning and neural networks for spore G/NG prediction and achieved AUC > 0.95. Nevertheless, they acknowledged the limited interpretability of their models and the relatively small dataset used for training.

In the context of microbiome-informed prediction, Ince et al. (2024) augmented metagenomic data using CTGAN to enhance model performance in classifying *Clostridium* growth patterns. Synthetic sample generation improved AUC from 0.79 to 0.94, but the base dataset remained small, impacting the overall reliability. Efforts to build more generalized models for microbial prediction using large-scale data were demonstrated by Hosoe et al. (2022) and Hiura et al. (2021), who utilized ensemble methods like XGBoost trained on ComBase data. Hosoe et al. (2022) modeled eight species under 5025 environmental conditions and achieved $R^2 = 0.76$, while Hiura et al. (2021) attained $R^2 = 0.75$ using 27,059 observations for *Listeria monocytogenes*. While both studies benefited from dataset scale and SHAP-based interpretability, Hosoe et al. (2022) specifically noted variability in model accuracy across microorganisms and food categories as a limitation.

Beyond growth modeling, ML has also been utilized in microbial source attribution. Munck et al. (2020) implemented logit boost and RF models to attribute human *Salmonella Typhimurium* cases to animal reservoirs using cgMLST profiles. The logit boost model yielded 92% accuracy and correctly attributed 81% of 118 human cases. However, their dataset was relatively small and imbalanced, and the authors noted reduced transparency in the interpretability of the model outputs.

Karlsen et al. (2021) explored the prediction of milk acidification rates (Vmax) of *Lactococcus lactis* using Random Forest models trained on 342 genome sequences with four genomic representations. The 9-mer model achieved the highest performance (PC = 0.85). Although the models revealed meaningful biological markers related to lactose metabolism and stress response, the study acknowledged the challenge of model interpretability and the limitations posed by the moderate dataset size.

Despite the growing application of machine learning in predictive microbiology, a number of studies still lack interpretability-enhancing tools such as SHAP or feature importance analysis, including (Al et al. 2024; Yildirim-Yalcin et al. 2025; Oliveira et al. 2024; Munck et al. 2020; Koyama et al. 2022), which limits the biological transparency and trustworthiness of their models. In addition, several investigations such as (Koyama et al. 2022; Oliveira et al. 2024; Tarlak 2024; Kuroda et al. 2019; Yildirim-Yalcin et al. 2025) are constrained by relatively small datasets, which can affect both the generalizability and robustness of the trained models.

Moreover, while *Lactococcus lactis* has been studied in the context of milk acidification kinetics (Karlsen et al. 2021), its abundance dynamics—particularly in response to microbial competition or environmental stress—remain unaddressed through ML-based modeling. No studies to date have examined the potential seasonal variation in microbial patterns by incorporating samples collected across different time points, which could reveal temporal signatures affecting microbial behavior. Finally, interspecies microbial relationships, which are critical for understanding microbial ecology in food systems, have not been thoroughly explored using explainable machine learning frameworks, leaving a significant gap in both mechanistic understanding and predictive insight. Lastly, given that data collection in this domain is often time-consuming and expensive, the underutilization of generative methods such as GAN-based data augmentation represents a significant research gap.

3 Material and Methods

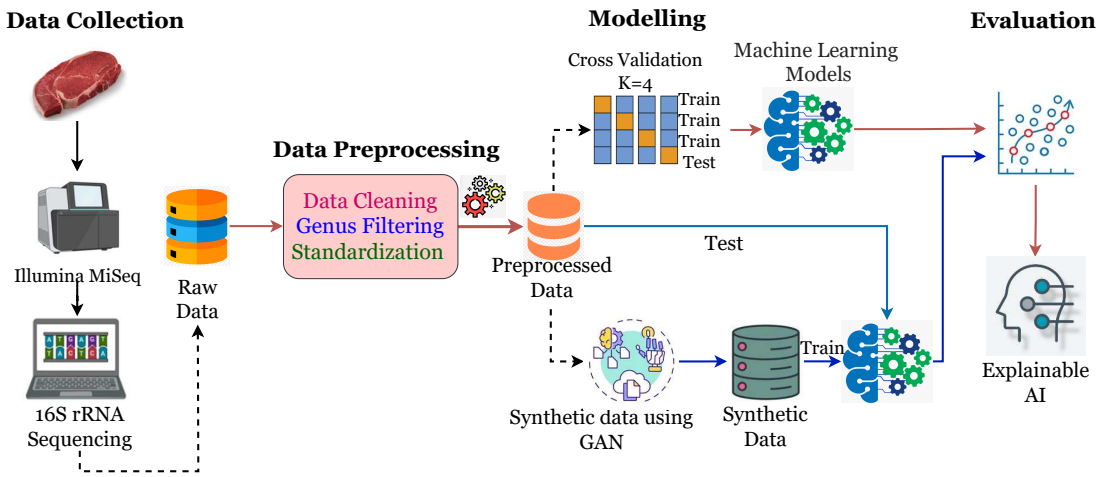


Fig. 1: The flow chart of the steps followed in the current study.

In this study, a multi-step analytical framework was employed to investigate microbial dynamics in fermented meat products. (1) Microbial community data were first collected through 16S rRNA gene sequencing using the Illumina MiSeq platform. (2) The resulting raw sequencing data were subjected to a comprehensive preprocessing pipeline that included data cleaning, genus-level filtering (where genera with over 80% of values below 0.1% relative abundance were removed), and standardization of features to ensure consistency in scale. (3) To augment the dataset and improve model robustness, synthetic microbial profiles were generated using a GAN. (4) These synthetic

data were combined with the original dataset to train supervised machine learning models, using cross-validation and test-train splitting strategies to evaluate predictive performance. (5) Finally, model evaluation was conducted using performance metrics such as R^2, MAE, and MSE, along with explainable AI techniques—specifically SHAP—to interpret feature contributions and identify key ecological and contextual drivers of *Lactococcus* abundance. The workflow diagram outlining these steps is illustrated in Figure 1.

3.1 Data Collection

Skin-packed rump steaks were shipped overnight to Campden BRI (Campden BRI) on kill day +21 (K+21) following processing at the packaging sites, representing retail day 0 (R+0). Upon arrival the next day (R+1), the samples were stored at $5°C \pm 1°C$. On R+12, the samples underwent temperature abuse at $22°C \pm 1°C$ for 2 hours, after which they were stored at $7°C \pm 1°C$ for the remainder of their shelf life to simulate consumer handling and storage in a domestic refrigerator. Samples were analyzed at R+1 (upon receipt), R+12 (before temperature abuse at 22°C), R+13, and R+18. One rump steak from each of the six trial carcasses was tested.

Rump steaks were homogenized at a 1:1 dilution using Maximum Recovery Diluent (MRD). The homogenized 1:1 sample was used for 16S rRNA metabarcoding. Further dilutions with MRD were prepared to create a 1:10 dilution and subsequent serial dilutions for culture-based microbiological testing.

3.2 16S rRNA Metabarcoding

For samples to undergo 16S rRNA metabarcoding and population analysis, bacterial genomic DNA was extracted from each sample using the QIAGEN Power food DNeasy kit following manufacturer's protocol. Samples were pooled together and processed on an Illumina MiSeq with cycling conditions: 300 x 2 bp per run. The sequencing data was analysed using the 16S Metagenomics (V1.1.0) application on the Illumina BaseSpace online web application. Amplicons were taxonomically ranked using the RefSeq RDP 16S (V3) database. To amplify the variable regions 3 and 4 of the 16S rRNA gene, the primers 341 F (TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTACGGGNGGCWGCAG) and 805 R (GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC) were used.

3.3 Dataset Overview

The dataset used in this study comprises microbiological and physicochemical data collected from 84 rump steak samples. A total of 734 bacterial species were identified based on abundance data. Samples were collected over seven different days (Day 4, 11, 21, 22, 33, 34, and 40) across two distinct seasons, Spring and Autumn. The dataset includes 736 features in total, of which 735 are numerical and represent bacterial abundance and other quantitative measurements, while 1 is categorical. The categorical feature Season indicates the season in which the sample was collected, with possible values being Spring and Autumn. This dataset offers a comprehensive profile of microbial diversity and environmental factors influencing spoilage over time and under various storage conditions. The dataset used in this study was provided as part of a collaborative effort with Campden BRI (Campden BRI).

3.4 Data Preprocessing

The initial dataset comprised microbial abundance information for 734 unique genera derived from 16S rRNA sequencing data. To reduce dimensionality and remove potentially non-informative features, a filtering step was applied. Specifically, genera that were present in less than 0.1% of the samples or appeared in more than 80% of the samples were excluded, as such features are either too sparse or overly dominant to contribute meaningful variance to the model. After this filtering process, a total of 31 genera remained. Categorical variables were processed accordingly. The *Season* feature, which includes categories such as spring and autumn, was transformed using one-hot encoding to convert the categorical labels into binary numerical features suitable for machine learning algorithms. Finally, to ensure that all input features contributed equally to the learning process and to improve model convergence, the dataset was standardized. This involved scaling all numerical features to have zero mean and unit variance.

3.5 GAN-Based Synthetic Dataset

In this study, the original dataset obtained from Campden BRI was expanded using the TVAESynthesizer (Tabular Variational Autoencoder) (SDV) to generate 1,000 synthetic data points. The TVAESynthesizer is a deep learning-based generative model designed to produce realistic synthetic data while preserving the statistical properties and underlying patterns of the original dataset (Khatiman et al. 2023). This approach enhances the dataset size and diversity, which can improve model generalization and increase the robustness of subsequent machine learning analyses. The generated synthetic data maintain the original data's structure, including the relationships between bacterial abundance, collection days, and seasonal variations (Kiran and Kumar 2024). This data augmentation strategy allows for more comprehensive training and testing of predictive models, thereby strengthening the validity and reliability of the research findings.

3.6 Machine Learning Models

ANN, Random Forest, k-Nearest Neighbors, and Gradient Boosting are widely used machine learning algorithms across various domains. This synthesis explores their methodologies, advantages, and applications.

3.6.1 Artificial Neural Networks

ANN are computational models inspired by the human brain's neural architecture. They consist of interconnected nodes or neurons organized into layers – input, hidden, and output layers. ANNs excel in capturing complex relationships within data and are particularly effective in tasks such as image and speech recognition, where deep learning can enhance their performance.(Zain et al. 2012).

3.6.2 Random Forest

Random Forest is an ensemble learning technique that constructs multiple decision trees during training and outputs the mode of their predictions (for classification) or the mean prediction (for regression). A significant advantage of RF is its robustness against overfitting, as it builds trees from random subsets of data and features, effectively reducing variance without greatly increasing bias (Breiman 2001)

3.6.3 k-Nearest Neighbors

k-Nearest Neighbors operates by identifying the 'k' closest samples in the feature space when predicting the output for a new instance. It's a non-parametric method, meaning it makes no explicit assumptions about the distribution of data. This algorithm shines in scenarios with well-defined class borders but struggles with high-dimensional data due to the curse of dimensionality, which can lead to decreased performance (Song et al. 2017)

3.6.4 Gradient Boosting

Gradient Boosting is another ensemble technique that builds models sequentially, where each new model attempts to correct the errors made by previous ones. It is characterized by its ability to reduce both bias and variance, making it a powerful tool for tasks requiring high predictive accuracy (Friedman 2001)

3.7 SHAP Analysis

SHAP analysis, or SHapley Additive exPlanations, emerges as a pivotal framework in the context of interpretable machine learning, grounded in cooperative game theory. This approach is designed to elucidate the contributions of individual features towards the predictive output of complex machine learning models, thereby enhancing their interpretability (Lundberg and Lee 2017). By calculating SHAP values, it quantifies the impact of each feature in a manner that maintains fairness across all contributing variables, effectively attributing the final prediction to their respective influences (Zargari Marandi 2024).

The SHAP value for a given feature i, denoted as ϕ_i, is defined by the following equation:

Table 1: Machine learning models hyperparameters used in this study.

Model	Hyperparameter	Value
ANN	Layers	5 (512, 512, 512, 512, 1)
	Activation	ReLU (output: Linear)
	Optimizer	Adam
	Learning rate	0.001
	Loss	Mean Squared Error
	Epochs	10
	Batch size	4
Random Forest	n_estimators	100
	criterion	squared_error
	max_depth	None
	min_samples_split	2
	min_samples_leaf	1
	max_features	auto
	random_state	42
Gradient Boosting	n_estimators	500
	learning_rate	0.01
	loss	squared_error
	max_depth	3
	subsample	0.7
	min_samples_split	2
	min_samples_leaf	5
	max_features	None
	random_state	4
KNN	n_neighbors	5
	weights	uniform
	algorithm	auto
	leaf_size	30
	metric	minkowski

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[f(S \cup \{i\}) - f(S) \right] \tag{1}$$

Where:

- ϕ_i: The SHAP value for feature i, representing its contribution to the prediction.
- N: The set of all input features.
- S: A subset of features excluding i.
- $f(S)$: The model's prediction for the subset S.

3.7.1 Cross-Validation

To evaluate the generalizability and performance of the machine learning models, we employed k-fold cross-validation with $k = 4$. In this approach, the dataset is divided into four equal-sized subsets (folds). During each iteration, three folds are used for training the model while the remaining fold is used for validation. This process is repeated four times such that each fold is used exactly once as the validation set (Nti et al. 2021).

The final evaluation metric is computed as the average of the performance scores obtained in each iteration. Let M_i denote the evaluation metric (e.g., mean squared error) for fold i, where $i = 1, 2, 3, 4$. The overall metric M is calculated as follows:

$$M = \frac{1}{k} \sum_{i=1}^{k} M_i \tag{2}$$

In our experiments, this procedure ensures a balanced and robust estimation of model performance, helping to mitigate issues such as overfitting and variance due to random train-test splits.

3.8 Model Evaluation Metrics

To evaluate the performance of the regression models, three key metrics were used: the coefficient of determination (R^2), Mean Squared Error (MSE), and Mean Absolute Error (MAE). These metrics provide a comprehensive understanding of the models' predictive accuracy and overall performance.

3.8.1 Coefficient of Determination (R^2)

The coefficient of determination (R^2) measures the proportion of the variance in the dependent variable that is predictable from the independent variables. An R^2 value close to 1 indicates that the model explains most of the variability in the target variable, whereas a value close to 0 suggests poor predictive performance (Di Bucchianico 2008). The R^2 value is calculated using the following equation:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

where:

- y_i = Actual value of the dependent variable for the i^{th} sample
- $\hat{y}_i$ = Predicted value of the dependent variable for the i^{th} sample
- $\bar{y}$ = Mean of the actual values
- n = Number of samples

3.8.2 Mean Squared Error

Mean Squared Error represents the average squared difference between the actual and predicted values. Lower MSE values indicate better predictive accuracy, as the model's predictions are closer to the true values (Wang and Bovik 2009). The MSE value is calculated using the following equation:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

3.8.3 Mean Absolute Error

Mean Absolute Error measures the average absolute difference between the actual and predicted values. Unlike MSE, MAE provides a more interpretable metric since it reflects the average error magnitude in the original units of the target variable (Hodson 2022). The MAE value is calculated using the following equation:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5}$$

4 Results

In this section, we present the results of our predictive modeling and microbial abundance analysis. We focus on evaluating the seasonal dynamics of key microbial genera, particularly *Lactococcus*, using both the original and GAN-generated synthetic datasets. The performance of different machine learning models is also assessed, followed by interpretability insights derived from SHAP analysis.

4.1 Comparative Analysis of Seasonal Variation in Microbial Genera Using Original and GAN-Based Synthetic Datasets

To evaluate the capacity of synthetic data in capturing biologically relevant patterns, we compared microbial genus-level abundances between spring and autumn across both the original and GAN-based synthetic datasets. Figure 2 displays the top 10 most abundant genera in each season. In both datasets, genera such as *Lactobacillus*,*Carnobacterium*, and *Leuconostoc* consistently ranked among the most dominant taxa. However, notable seasonal shifts in relative abundance were observed, with some genera showing substantial increases in autumn compared to spring.

Focusing on *Lactococcus*, a genus exhibiting pronounced seasonal variability, we observed a statistically significant difference in abundance between seasons in both datasets. As shown in Figure

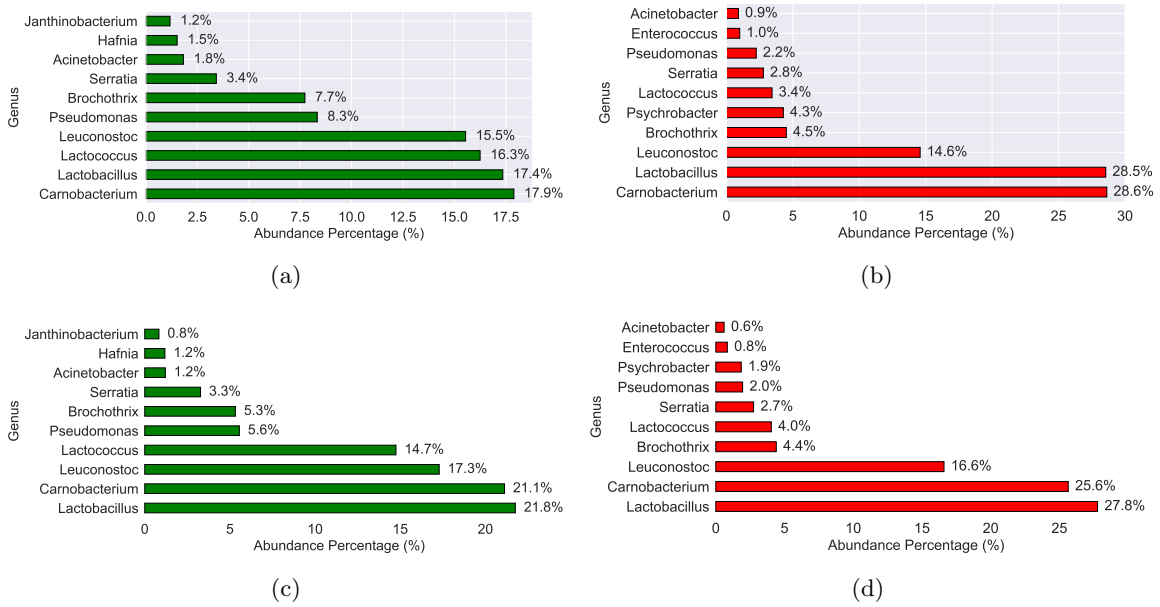


Fig. 2: Top 10 most abundant genera in samples collected in different seasons: (a) Top 10 genera in terms of abundance percentage in samples collected in autumn (original dataset). (b) Top 10 genera in terms of abundance percentage in samples collected in spring (original dataset). (c) Top 10 genera in terms of abundance percentage in samples collected in autumn (GAN-based synthetic dataset). (d) Top 10 genera in terms of abundance percentage in samples collected in spring (GAN-based synthetic dataset).

3, the original dataset yielded a p-value of 0.0001 (Figure 3a), whereas the GAN-based synthetic dataset produced an even more significant result (p = 8.60e–40, Figure 3b). In both cases, *Lactococcus* abundance was markedly higher in autumn than in spring, suggesting a robust seasonal signal that is well preserved in the synthetic data.

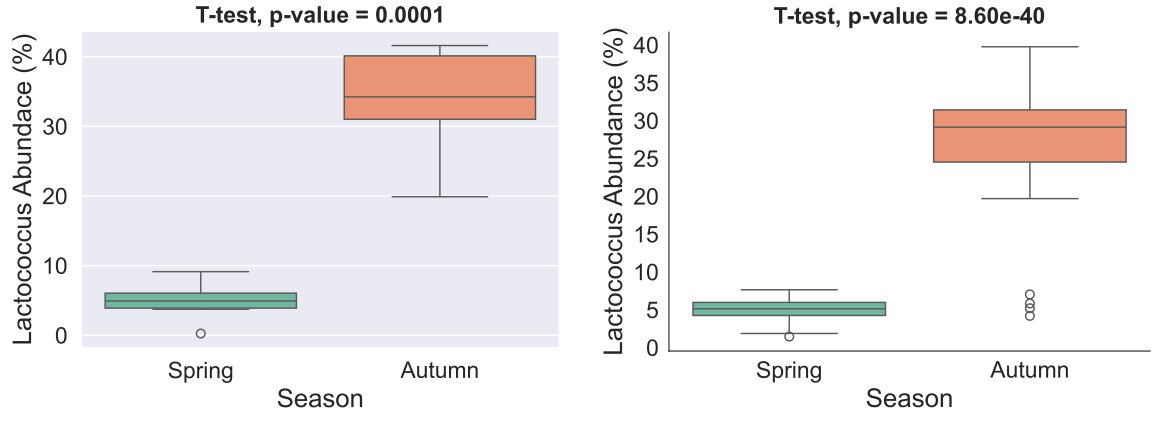


Fig. 3: Box plot showing the abundance of Lactococcus in samples collected during spring and autumn. (a) Original dataset (T-test, p-value = 0.0001); (b) GAN-based synthetic dataset (T-test, p-value = 8.60e-40). The low p-values in both datasets indicate a statistically significant difference between the two seasons.

To further dissect these seasonal trends, we analyzed *Lactococcus* abundance across individual sampling days (Figures 4 and 5). In the original dataset (Figure 4), statistically significant differences emerged from Day 21 onward (e.g., Day 21: p = 0.003739; Day 33: p = 1.622e–05), with autumn samples consistently exhibiting higher *Lactococcus* abundance. The GAN-based synthetic dataset

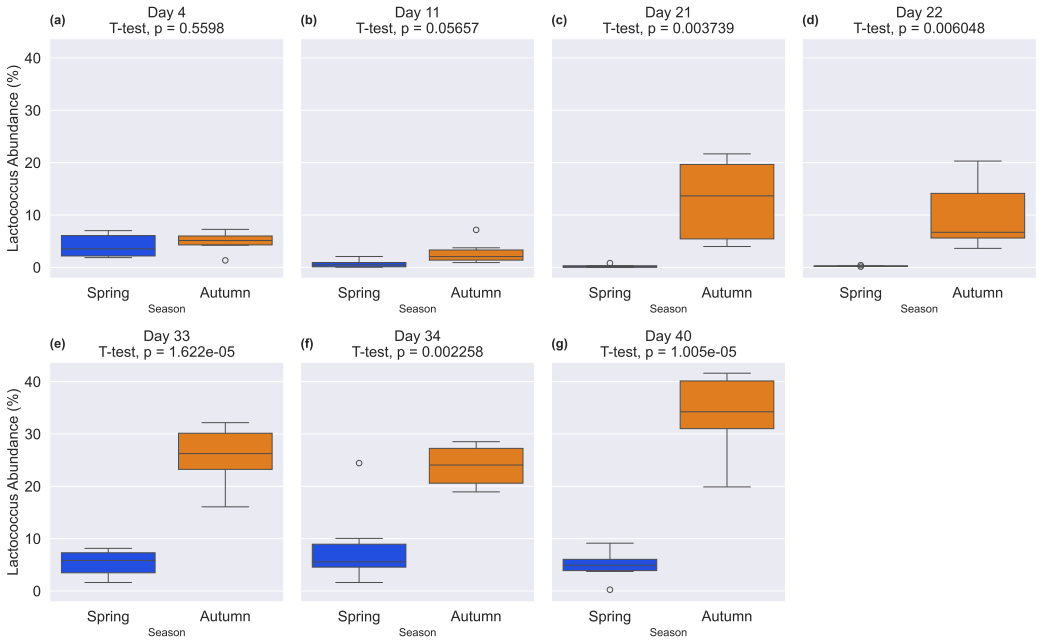


Fig. 4: Box plots showing the abundance of *Lactococcus* in samples collected during spring and autumn on different days for the original dataset. The p-values from t-tests indicate the statistical significance of the differences in abundance between seasons: (a) Day 4, (b) Day 11, (c) Day 21, (d) Day 22, (e) Day 33, (f) Day 34, (g) Day 40. Blue boxes represent spring samples, while orange boxes represent autumn samples.

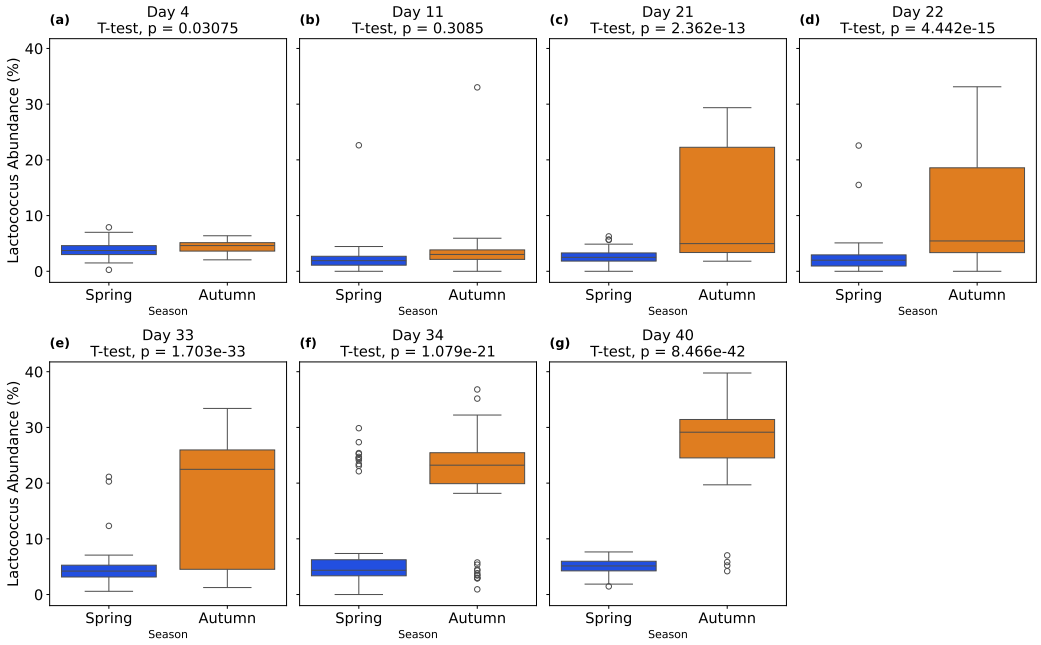


Fig. 5: Box plots showing the abundance of *Lactococcus* in spring and autumn samples on different days for the GAN-based synthetic dataset. The p-values from t-tests indicate the statistical significance of the differences in abundance between seasons: (a) Day 4, (b) Day 11, (c) Day 21, (d) Day 22, (e) Day 33, (f) Day 34, (g) Day 40. Blue boxes represent spring samples, while orange boxes represent autumn samples.

mirrored these patterns but exhibited even stronger statistical significance at earlier time points (e.g., Day 21: p = 2.362e–13; Day 33: p = 1.703e–33), indicating enhanced sensitivity to temporal variation.

These results collectively underscore the capacity of GAN-generated synthetic datasets to not only replicate the overarching seasonal patterns evident in the original data but also to amplify nuanced biological signals that may be less discernible in empirical observations. The high degree of concordance between the real and synthetic datasets substantiates the applicability of GAN-based methodologies in microbiome research, particularly in contexts requiring data augmentation or the mitigation of sampling-related limitations. Furthermore, the faithful retention of ecologically relevant dynamics—such as the seasonal variability of *Lactococcus*—demonstrates that synthetic data can function as biologically credible surrogates for empirical datasets, thereby offering a robust framework for downstream ecological and functional inference.

4.2 Correlation Analysis Between Features and *Lactococcus* Abundance

Table 2: Top 10 absolute Spearman correlations with corresponding p-values for original and GAN-based synthetic datasets.

Original Dataset			**GAN-based Synthetic Dataset**		
Feature	**Spearman Correlation**	***p*-value**	**Feature**	**Spearman Correlation**	***p*-value**
Hafnia	0.67	4.64×10^{-12}	*Hafnia*	0.547240	3.39×10^{-79}
Buttiauxella	0.56	3.52×10^{-8}	*Carnobacterium*	-0.542577	1.26×10^{-77}
Day	0.48	4.81×10^{-6}	*Bacillus*	-0.509052	4.95×10^{-67}
Leuconostoc	0.43	4.09×10^{-5}	*Staphylococcus*	-0.487301	9.15×10^{-61}
Weissella	0.37	6.32×10^{-4}	*Rubrobacter*	-0.434959	2.05×10^{-47}
Carnobacterium	-0.57	1.35×10^{-8}	*Janthinobacterium*	-0.434317	2.90×10^{-47}
Psychrobacter	-0.58	7.66×10^{-9}	*Weissella*	0.429350	4.08×10^{-46}
Bacillus	-0.62	4.50×10^{-10}	*Brochothrix*	-0.419887	5.55×10^{-44}
Exiguobacterium	-0.51	8.67×10^{-7}	Day	0.457883	5.62×10^{-53}
Staphylococcus	-0.47	7.86×10^{-6}	*Flavobacterium*	-0.372644	2.69×10^{-34}

To further evaluate the factors potentially associated with *Lactococcus* abundance, we conducted Spearman correlation analyses on both the original and GAN-based synthetic datasets. Table 2 summarizes the top 10 features exhibiting the strongest absolute correlation values, along with their corresponding p-values.

In the original dataset, *Hafnia* exhibited the highest positive correlation with *Lactococcus* abundance ($\rho = 0.67$, $p = 4.64 \times 10^{-12}$), followed by *Buttiauxella* and the variable *Day*, suggesting a temporal component to *Lactococcus* dynamics. Notably, several genera such as *Carnobacterium*, *Psychrobacter*, and *Bacillus* showed significant negative correlations ($\rho = -0.57$ to -0.62), indicating potential antagonistic or mutually exclusive relationships in microbial community structure.

The GAN-based synthetic dataset yielded a comparable correlation structure, with *Hafnia* again demonstrating a strong positive correlation ($\rho = 0.547$, $p = 3.39 \times 10^{-79}$), consistent with the original data. Similarly, *Carnobacterium* and *Bacillus* showed robust negative associations with *Lactococcus* abundance ($\rho = -0.543$ and -0.509, respectively), reinforcing the biological plausibility of the synthetic model's outputs. Importantly, the synthetic data also captured additional genera—such as *Janthinobacterium*, and *Flavobacterium*—exhibiting significant correlations, potentially reflecting an expanded or nuanced representation of community interactions.

The high concordance between key correlates in the real and synthetic datasets substantiates the capacity of GAN-based models to preserve underlying microbial association patterns. Moreover, the extremely low p-values observed in the synthetic dataset indicate a high level of statistical confidence, further validating the fidelity of synthetic data in capturing ecologically meaningful relationships.

4.3 Performance Evaluation of Machine Learning Models on the Original Dataset

Table 3 provides a comparative performance analysis of four widely used machine learning algorithms—ANN, Random Forest, KNN, and Gradient Boosting—applied to the original dataset. The

Table 3: Performance comparison of machine learning models using cross-validation on the original dataset. Reported metrics include the coefficient of determination (R^2), MSE, and MAE.

Algorithms	R^2 (Mean ± Std)	MSE (Mean ± Std)	MAE (Mean ± Std)
ANN	0.87 ± 0.04	15.86 ± 4.24	2.70 ± 0.44
Random Forest	0.74 ± 0.09	31.12 ± 9.10	3.99 ± 0.52
KNN	0.81 ± 0.03	23.69 ± 4.03	3.31 ± 0.31
Gradient Boosting	0.77 ± 0.10	27.73 ± 10.04	3.79 ± 0.50

evaluation was performed using three regression metrics: coefficient of determination (R^2), MSE, and MAE, with each metric reported as the mean ± standard deviation over 4-fold cross-validation.

Among the evaluated models, the ANN consistently outperformed the others across all metrics. It achieved the highest R^2 score (0.87 ± 0.04), indicating that approximately 87% of the variance in the target variable could be explained by the model. Moreover, ANN yielded the lowest MSE (15.86 ± 4.24) and MAE (2.70 ± 0.44), reflecting superior predictive accuracy and minimal deviation from actual observations. These results underscore the ANN's capability to capture complex, non-linear interactions within microbiome data. Gradient Boosting and KNN followed with comparable R^2 scores (0.87 and 0.81, respectively), although KNN had a lower MSE (23.69 ± 4.03) compared to Gradient Boosting (27.73 ± 10.04), suggesting better average error performance despite its simpler architecture. Random Forest demonstrated the lowest R^2 (0.74 ± 0.09) and the highest MSE (31.12 ± 9.10), indicating suboptimal performance relative to the other models. Additionally, Random Forest and Gradient Boosting both exhibited relatively higher standard deviations across folds, implying less stable generalization performance.

These results are further supported by the scatter plots in Figure 6, which visualize the relationship between predicted and actual values for each model. The ANN model (Figure 6a) shows points tightly clustered around the perfect fit line ($y = x$), confirming its high accuracy. In contrast, the predictions from the Random Forest model (Figure 6b) display greater dispersion from the ideal line, consistent with its lower R^2 score. Similarly, Gradient Boosting and KNN models (Figures 6c and 6d) exhibit intermediate levels of predictive alignment.

Complementary to this, Figure 7 presents the predicted and true Lactococcus abundance values across all test samples. Again, the ANN model (Figure 7a) demonstrates the closest alignment between predicted (red) and true (blue) values, with less deviation across sample indices. Random Forest (Figure 7b) shows more pronounced discrepancies, particularly in underestimating or overestimating peak abundance values. Gradient Boosting and KNN (Figures 7c and 7d) partially capture the overall trend but miss certain local fluctuations.

The consistently strong results of ANN highlight its suitability for regression tasks involving microbiome data, where variable interdependencies and non-linear relationships are prevalent. The observed variation in model performance also reinforces the need for careful model selection based on both accuracy and consistency. These findings provide a robust baseline for future predictive modeling efforts using microbiome features and underscore the feasibility of achieving high prediction accuracy with appropriate machine learning frameworks.

4.4 Performance of Models Trained on GAN-Based Synthetic Data

Table 4 illustrates the generalization capabilities of machine learning models trained exclusively on the GAN-based synthetic dataset and evaluated on the original dataset. The reported metrics coefficient of determination (R^2), MSE, and MAE—provide insight into the extent to which synthetic data can effectively support predictive modeling tasks in real-world contexts.

Remarkably, all four models demonstrated substantial predictive performance when applied to the original dataset, despite being trained solely on synthetic data. The ANN model once again emerged as the top performer, achieving the highest R^2 (0.89), the lowest MSE (13.51), and the lowest MAE (2.54). These results not only mirror the model's success on real-data training (as observed in Table 4) but also indicate a strong alignment between the synthetic and empirical data distributions. Gradient

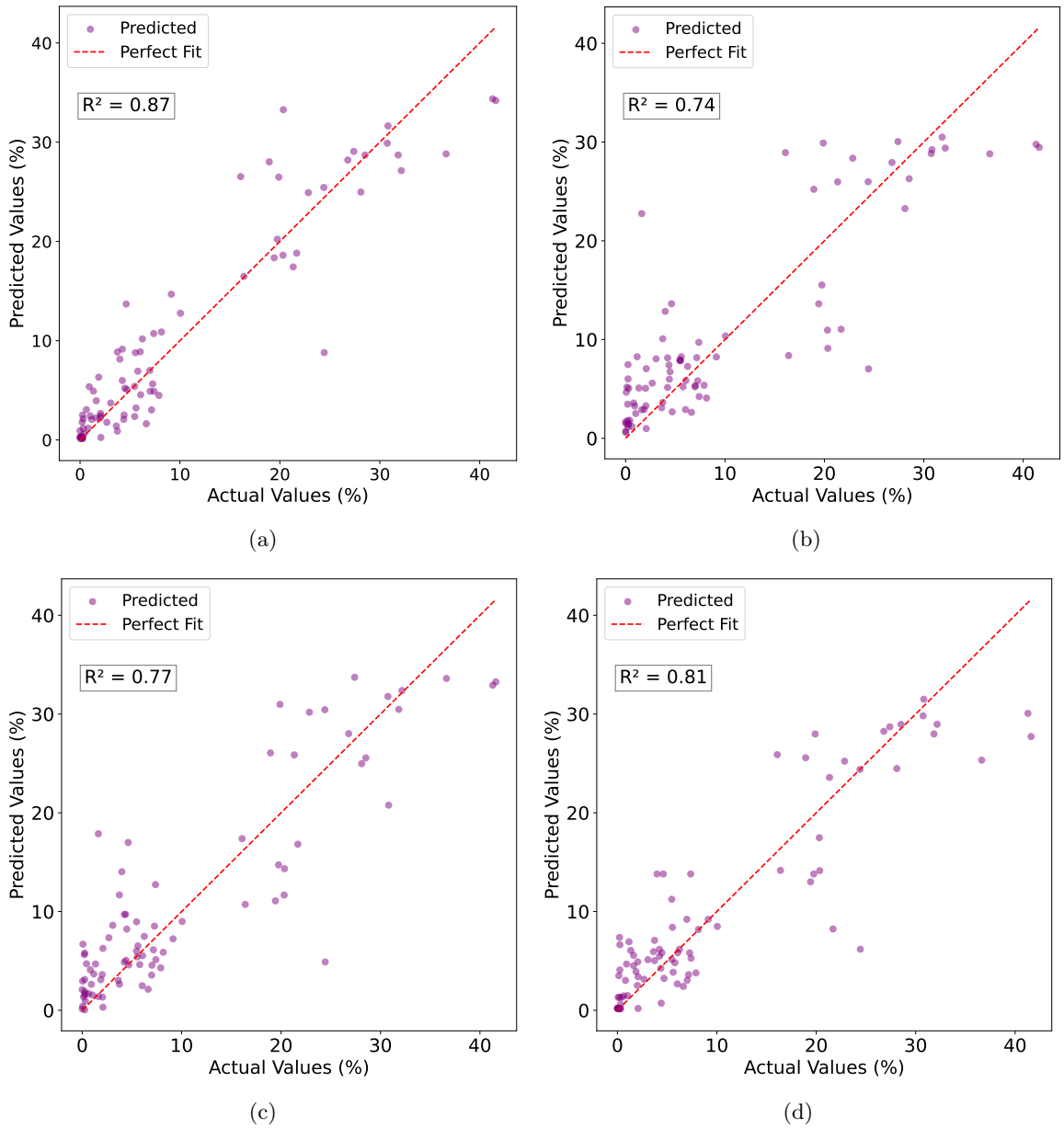


Fig. 6: Scatter plots of predicted versus actual values for different regression models on the original dataset: (a) ANN model, (b) Random Forest model, (c) Gradient Boosting Regression model, (d) KNN model. The red dashed line represents the perfect fit line ($y = x$), and R^2 values indicate the goodness of fit for each model.

Table 4: Performance of machine learning models trained on the GAN-based synthetic dataset and evaluated on the original dataset. Reported metrics include the coefficient of determination (R^2), MSE, and MAE.

Model	R^2	**MSE**	**MAE**
ANN	0.89	13.51	2.54
Random Forest	0.85	19.03	3.13
Gradient Boosting	0.87	15.95	2.93
KNN	0.81	23.05	3.07

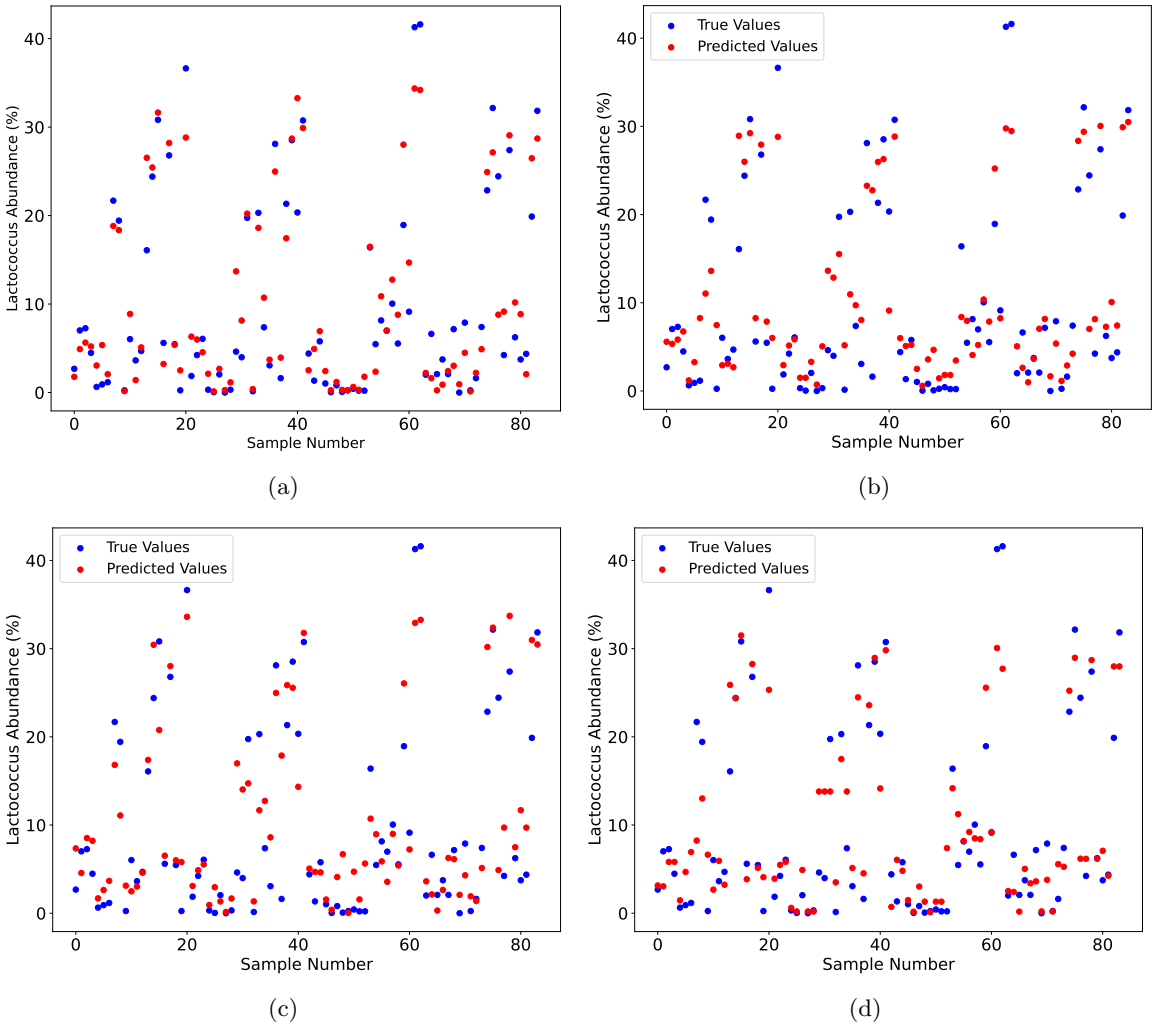


Fig. 7: Comparison of true and predicted *Lactococcus* abundance values across different models: (a) ANN model, (b) Random Forest model, (c) Gradient Boosting Regression model, (d) KNN model. Blue dots represent true values, while red dots represent predicted values.

Boosting and Random Forest also achieved competitive R^2 values (0.87 and 0.85, respectively), with relatively low error margins, further validating the biological plausibility of the GAN-generated dataset. The KNN model, while exhibiting slightly reduced accuracy ($R^2 = 0.81$), still maintained reasonable generalization, suggesting that even simpler algorithms can extract meaningful patterns from synthetic inputs.

To complement the quantitative results in Table 4, Figure 8 presents scatter plots of predicted versus actual values for each model trained on the GAN-based synthetic dataset and tested on the original data. The ANN model (Figure 8a) shows the closest alignment with the perfect prediction line ($y = x$), achieving the highest R^2 value (0.89), indicating strong generalization. Gradient Boosting (0.87) and Random Forest (0.85) follow closely, while KNN (0.81) exhibits more dispersion, particularly at higher values. These results visually confirm the effectiveness of the synthetic data in supporting real-world prediction tasks, particularly for the ANN model.

These findings underscore the utility of GAN-based synthetic data in enabling model training when real data is scarce, sensitive, or imbalanced. The ability of models trained on synthetic samples to generalize effectively to actual data highlights the potential of generative approaches in microbiome research and beyond. This also creates opportunities to use synthetic data for model pretraining, transfer learning, and data augmentation, which can help improve the reliability and performance of predictive models in complex biological applications.

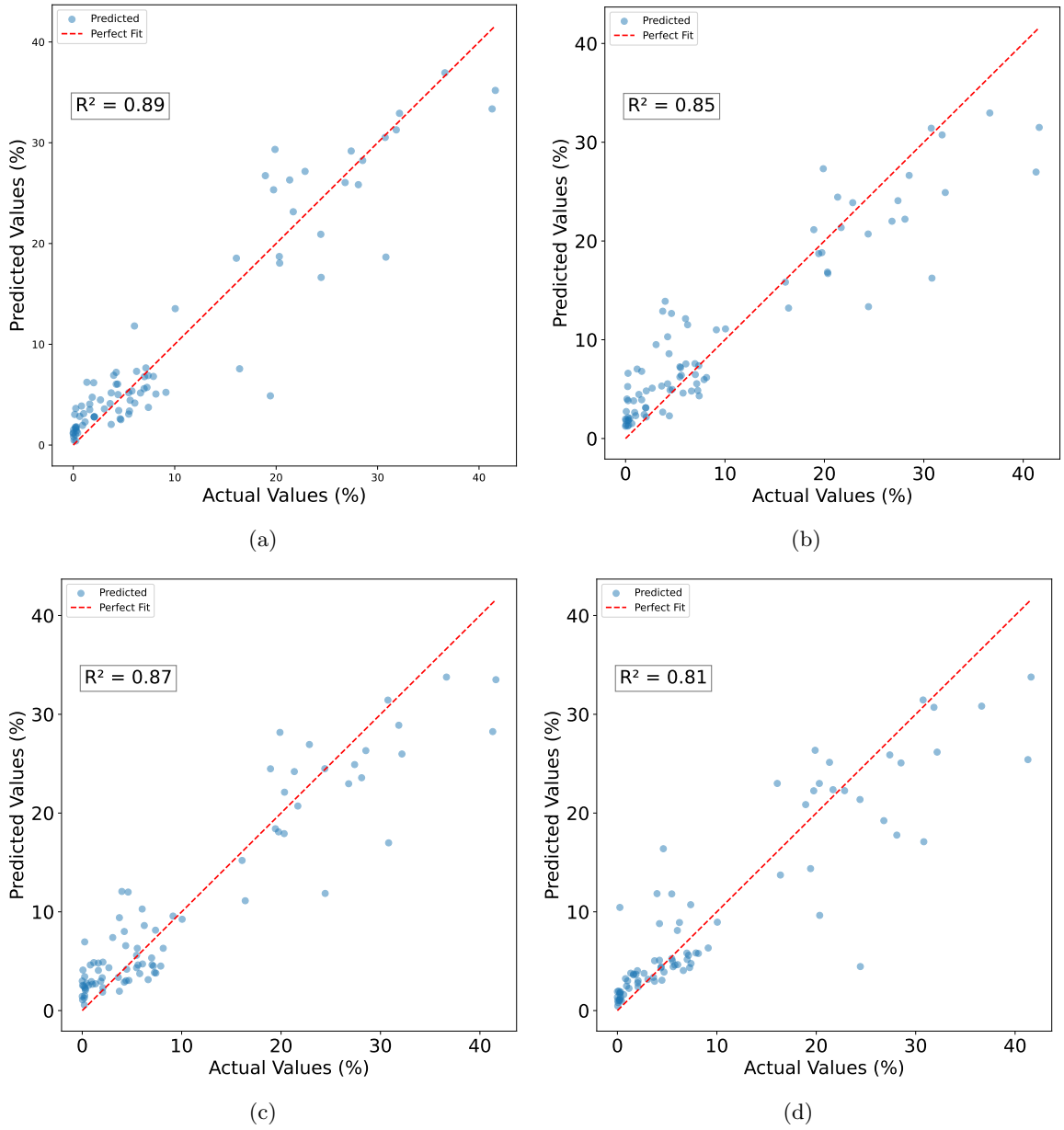


Fig. 8: Scatter plots of predicted versus actual values for different regression models on the GAN-based synthetic dataset: (a) ANN model, (b) Random Forest model, (c) Gradient Boosting Regression model, (d) KNN model. The red dashed line represents the perfect fit line ($y = x$), and R^2 values indicate the goodness of fit for each model.

4.5 Explaining Model Predictions through SHAP Value Results

The interpretability of the artificial neural network model developed for predicting *Lactococcus* abundance was evaluated using SHAP analysis based on both the original dataset and the GAN-generated synthetic dataset as presented in Figure 9. Figures 9a and 9b correspond to the SHAP summary plot and the mean absolute SHAP value plot for the original data respectively. In the summary plot each point represents the SHAP value for a single observation where the position on the x-axis reflects the impact of the corresponding feature on the model output. Points located on the right-hand side contribute positively to the predicted *Lactococcus* abundance whereas those on the left have a negative effect. The color of the points indicates the magnitude of the feature values with red denoting high values and blue representing low values

Upon examination of Figure 9a, *Lactobacillus* emerges as the most influential variable in predicting the abundance of *Lactococcus*. This is followed by *Season*, *Carnobacterium*, *Lactigenium*, and *Day*

(storage duration). The model outputs indicate that high levels of *Lactobacillus* are negatively associated with *Lactococcus* levels. This suggests a potential competitive relationship between the two species, wherein the dominance of *Lactobacillus* within the microbial composition may inhibit the growth of *Lactococcus* through mechanisms such as nutrient competition or metabolite production. Similarly, both *Carnobacterium* and *Lactigenium* exhibit negative associations with *Lactococcus*, further indicating the presence of comparable microbial interactions. In terms of the *Season* variable, the spring season appears to negatively influence *Lactococcus* abundance. This finding implies that seasonal environmental factors (e.g., temperature, humidity, surrounding microbiota) or variations in production and storage conditions may play a critical role in shaping the growth dynamics of this bacterial species. Lastly, the increase in the *Day* variable, which refers to the duration of storage, is positively correlated with *Lactococcus* levels. This observation suggests that *Lactococcus* may proliferate over time or that its relative abundance increases as other competing species decline during extended storage. Collectively, these findings provide important insights into the ecological dynamics of *Lactococcus* within microbial communities and underscore its relevance in maintaining microbial balance in food products.

Figures 9c and 9d display the SHAP analysis for the GAN-based synthetic dataset. In Figure 9c, the model was trained using synthetic data generated by a Generative Adversarial Network (GAN) and subsequently tested on the original dataset to assess the transferability of learned patterns. In this case, *Season* again emerged as the most important predictor of *Lactococcus* abundance, followed by *Weissella*, *Carnobacterium*, and *Hafnia*. Notably, high levels of *Weissella* and *Hafnia* were positively associated with increased *Lactococcus* levels, which may indicate synergistic or co-occurring microbial relationships. Conversely, *Carnobacterium* exhibited a negative association with *Lactococcus*, consistent with the pattern observed in Figure 9a and reinforcing the idea of microbial competition. Although the *Day* variable was ranked lowest in importance, it still demonstrated a positive influence on *Lactococcus* levels, suggesting a consistent trend where longer storage durations favor the presence or growth of this bacterial genus. Overall, the comparison of both models—one trained on real data and the other on GAN-generated synthetic data—provides valuable insights into the robustness of predictive features and the underlying ecological relationships within the microbiome.

5 Discussion

This study highlights the effectiveness of combining real microbiome data with GAN-generated synthetic data to model the seasonal variation of *Lactococcus* abundance in meat products. The results indicate that synthetic data not only preserve biological patterns observed in real datasets but also enhance the performance and generalizability of machine learning models. This dual advantage addresses two major challenges in microbiome-based predictive modeling: limited sample sizes and model interpretability.

Statistical comparisons between the original and GAN-based datasets revealed highly consistent trends in seasonal dynamics. Specifically, *Lactococcus* abundance was significantly higher in autumn than in spring in both datasets. The original data showed a statistically significant difference with a p-value of 0.0001, whereas the synthetic data presented an even stronger result ($p = 8.60e{-}40$), suggesting enhanced sensitivity to seasonal effects. These consistent findings across multiple time points, especially from day 21 onward, demonstrate the ability of synthetic data to preserve fine-grained biological variation and temporal patterns.

Further analysis revealed a high level of agreement in microbial correlations between the real and synthetic datasets. For instance, *Hafnia* was positively associated with *Lactococcus* abundance in both datasets (Spearman's $\rho = 0.67$ for real data and $\rho = 0.547$ for synthetic data), while *Carnobacterium* and *Bacillus* showed negative correlations. These relationships suggest competitive or mutually exclusive ecological interactions, and their preservation in synthetic data validates the biological credibility of the generative approach. Moreover, the synthetic dataset revealed additional correlated genera, such as *Janthinobacterium* and *Flavobacterium*, indicating that GAN-generated data may help uncover subtle relationships that are not easily detectable in small datasets due to limited statistical power.

From a predictive modeling perspective, the ANN consistently outperformed other machine learning models, both when trained on real and synthetic data. When trained on the original dataset, the ANN achieved an R^2 score of 0.87. Remarkably, training the ANN exclusively on synthetic data and testing it on real samples resulted in an even higher R^2 score of 0.89. This improvement reflects

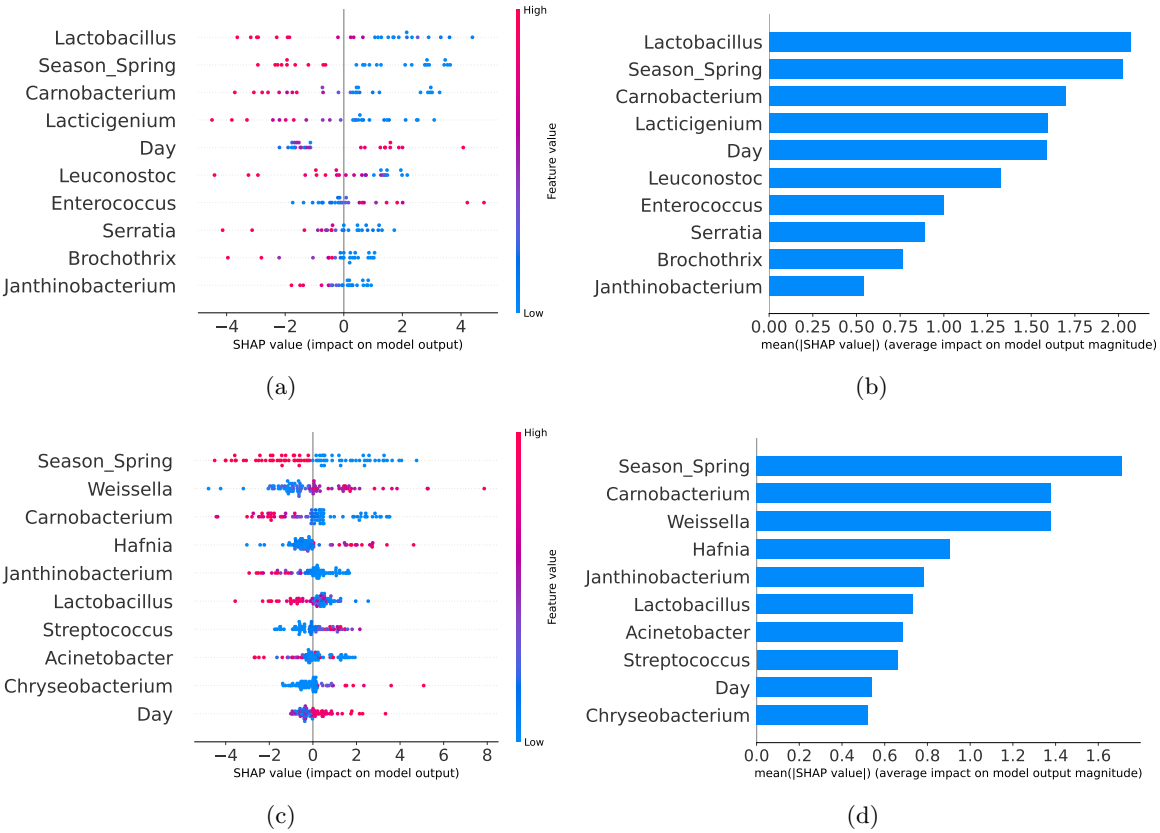


Fig. 9: SHAP summary plots and feature importance rankings for original and GAN-based synthetic datasets. (a) SHAP summary plot for the original dataset. (b) Mean absolute SHAP value plot showing the average impact of each feature on the model output for the original dataset. (c) SHAP summary plot for the GAN-based synthetic dataset. (d) Mean absolute SHAP value plot showing the average impact of each feature on the model output for the GAN-based synthetic dataset.

the high quality and utility of the GAN-generated data, which allowed the model to generalize well to previously unseen real-world data. Other models, including Gradient Boosting and Random Forest, also benefited from training on synthetic data, with R^2 values increasing from 0.77 and 0.74 to 0.87 and 0.85, respectively. Previous studies conducted in various domains using techniques such as GAN-based synthetic data augmentation (Arora and Arora 2022; Mert 2023) have demonstrated improvements in model accuracy and stability; these enhancements further support the argument that synthetic data augmentation can positively influence model performance.

To better understand the decision-making process of the models, SHAP analysis was applied. SHAP-based explainability enhances the interpretability of the model outputs, allowing stakeholders to identify which features most influence microbial trends (Hosoe et al. 2022). This explainability technique quantified the contribution of each feature to the prediction of *Lactococcus* abundance, offering transparency and interpretability for otherwise complex models. Across both datasets, features such as *Lactobacillus*, *Carnobacterium*, *Weissella*, *Hafnia*, and the `Season` variable were identified as the most influential predictors. Notably, *Lactobacillus* had a strong negative effect on *Lactococcus*, suggesting a possible competitive interaction. Similarly, *Carnobacterium* and *Lacticigenium* showed negative contributions, while *Hafnia* and *Weissella* were positively associated, especially in the synthetic dataset. These results reflect a mixture of competitive and cooperative relationships within the microbial community, aligning well with ecological theory. The SHAP analysis also revealed the importance of seasonal and temporal variables. The *Season* feature, particularly spring, was consistently linked to lower *Lactococcus* levels, whereas higher levels were observed in autumn. This suggests that environmental factors such as temperature and humidity, which vary seasonally, may impact microbial community structure and dynamics. For instance, Yap et al. (2024) observed higher *Lactococcus* activity during the summer months in raw milk samples, in contrast to the present

study, which identified increased *Lactococcus* abundance in meat products during autumn. Additionally, the storage duration, captured by the *Day* variable, was positively correlated with *Lactococcus* abundance. This trend implies that either *Lactococcus* becomes more competitive over time or its relative abundance increases as other microbial populations decline during extended storage.

Importantly, the SHAP feature rankings obtained from models trained on synthetic data closely matched those from real data, supporting the idea that GAN-based data can be used not only for improving prediction performance but also for extracting reliable and interpretable biological insights. This reinforces the potential of synthetic data in microbiome research, where interpretability and biological relevance are essential, particularly in food safety applications.

However, the study is not without limitations. One of the limitations of this study is that it was conducted solely on rump steak. This restricts the generalizability of the findings and provides limited insight into the applicability of the method to other food products. Additionally, the study focused on two seasons (spring and autumn), thus excluding potential variations during summer and winter that might also influence microbial behavior.

In conclusion, this study illustrates that combining synthetic data generation with interpretable machine learning can significantly enhance the modeling of microbial dynamics in food systems. GAN-generated data can serve as a reliable substitute or supplement for real datasets, especially when sample collection is limited or costly. SHAP analysis adds value by identifying the key drivers behind model predictions, bridging the gap between predictive performance and scientific understanding.

5.1 Comparing previous research

A wide range of studies have investigated the use of artificial intelligence methods in food safety assessment, leading to notable advancements in predictive performance and the overall reliability of such systems. However, critical challenges persist in terms of model interpretability, generalizability, and the integration of emerging technologies such as XAI and generative models. The comparative analysis provided in this section, drawing on previous studies outlined in Table 5, supports this discussion.

In terms of predictive accuracy, numerous studies have demonstrated success in modeling microbial behavior under diverse food system conditions. For instance, Koyama et al. (2022) reported an R^2 of 0.93 using support vector regression to model *E. coli* growth, while Al et al. (2024) achieved an R^2 of 0.98 employing Random Forest models on beef datasets. Similarly, Yücel and Tarlak (2023) attained $R^2 = 0.949$ with Random Forest Regression on the expansive ComBase dataset. Despite these impressive outcomes, many of the referenced studies utilized relatively constrained datasets and focused primarily on spoilage organisms, often overlooking beneficial microbial dynamics. Distinct from these efforts, the present study centers on *Lactococcus lactis*—a beneficial bacterium integral to natural food preservation due to its nisin production. ANN trained on synthetic data and tested on real samples achieved an R^2 of 0.89, indicating high performance even under cross-domain generalization. This positions the current research as a rare yet valuable contribution that applies high-performing ML models to the study of bioprotective rather than spoilage-associated bacteria.

Another key advancement in this work lies in its integration of explainable artificial intelligence, specifically SHAP values, to improve model transparency. Prior studies—such as those by Oliveira et al. (2024); Yildirim-Yalcin et al. (2025); Kuroda et al. (2019); Al et al. (2024); Munck et al. (2020); Koyama et al. (2022)—have largely emphasized predictive performance, often at the cost of interpretability. In contrast, our framework provides insights into the contribution of individual features to the model's predictions. Important factors such as *Lactobacillus*, *Carnobacterium*, and seasonal variables emerged as key determinants of *Lactococcus* abundance. This interpretable modeling approach enhances both scientific understanding and stakeholder confidence, which is especially critical in regulatory contexts where trust in AI systems is paramount.

Furthermore, the present study offers a novel perspective by examining interspecies microbial interactions—an area often neglected in previous literature. Many earlier models operate under the assumption that microbial species behave independently, thereby missing ecologically significant dynamics (Tarlak and Yücel 2023; Al et al. 2024; Hiura et al. 2021; Koyama et al. 2022). Our results, however, highlight both competitive and synergistic relationships: negative correlations were observed between *Carnobacterium* and *Lactococcus*, while positive associations with *Lactobacillus* suggest potential cooperation. These patterns were consistently supported by both SHAP values and Spearman correlation analysis, providing biologically meaningful interpretations of microbial community structure.

Lastly, data scarcity remains a significant limitation across microbiome modeling efforts (Papoutsoglou et al. 2023). Although recent studies have explored data synthesis to counter this issue—such as Ince et al. (2024), who applied CTGAN for improved classification—such approaches are infrequently applied to abundance prediction tasks. Building on this direction, our study utilized a Tabular Variational Autoencoder-based GAN to generate 1,000 synthetic samples that faithfully reproduced the statistical and ecological characteristics of the original dataset. The ANN trained exclusively on these synthetic samples not only generalized effectively to real-world data, but also outperformed other models, achieving the highest R^2 of 0.89. These findings highlight the considerable promise of generative methods in bolstering microbiome analyses, particularly when data availability is limited.

Table 5: Comparison of previous works on ML-based microbial prediction in food systems.

Reference	Techniques	Problem Addressed	Key Findings
Koyama et al. (2022)	SVR, ETR, GPR	Growth prediction of *E. coli* O157	SVR: $R^2 = 0.93$
Al et al. (2024)	RF, ANN, SVR, MLR	*E. coli* O157:H7 in raw beef	RF: $R^2 = 0.98$
Yücel and Tarlak (2023)	RFR, DTR, GAMR	Growth modeling of pathogens in beef	RFR: $R^2 = 0.949$
Tarlak (2024)	GPR, SVR, RFR	*Pseudomonas* in culture media	GPR; $R^2_{\text{adj}} = 0.959$
Tarlak and Yücel (2023)	RFR, GPR, DTR	Spoilage prediction in food products	RFR: $R^2 = 0.913$
Yildirim-Yalcin et al. (2025)	GPR, SVR, DTR	Mesophilic bacteria in spinach	GPR: $R^2 = 0.9888$
Oliveira et al. (2024)	Logistic Regression	G/NG modeling of *C. sporogenes*	62.8% accuracy
Sun et al. (2021)	Logistic Regression	G/NG of acid-tolerant *Bacillus*	c − value > 0.98
Kuroda et al. (2019)	Deep Learning	G/NG prediction of *Bacillus* spp.	AUC > 0.95
Ince et al. (2024)	CTGAN + RF	Synthetic data for *Clostridium*	AUC = 0.94
Hosoe et al. (2022)	XGBoost + SHAP	8 species across 15 food types	$R^2 = 0.76$
Hiura et al. (2021)	XGBoost	*L. monocytogenes* dynamics	$R^2 = 0.75$
Munck et al. (2020)	Logit Boost, RF	Source attribution of *Salmonella*	92% accuracy
Karlsen et al. (2021)	RF + Genomic Features	Vmax in *L. lactis* strains	$r = 0.85$

6 Conclusion

This study introduces an interpretable machine learning approach to predict the relative abundance of *Lactococcus* in a meat matrix, using both real and GAN-generated synthetic metagenomic data. The integration of explainable AI techniques with generative modeling provides a novel and effective framework for understanding microbial behavior under varying storage and environmental conditions. By enriching the training data with synthetic profiles, the models developed in this work were able to capture complex patterns and relationships that would otherwise remain hidden due to limited sample sizes.

The results demonstrated that artificial neural networks were particularly effective in modeling the microbial composition, capturing subtle ecological interactions between *Lactococcus* and other bacterial genera. Explainability tools such as SHAP enabled the identification of key contributors to prediction outcomes, highlighting the roles of microbial co-occurrence, storage time, and seasonal variability. This not only improves model transparency but also contributes to our biological understanding of how certain beneficial bacteria behave within food systems.

The use of synthetic data to simulate microbial profiles underrepresented in experimental datasets emerged as a promising strategy to address common challenges in microbiome research, such as cost constraints and limited experimental scope. Furthermore, the consistency observed between results obtained from real and synthetic datasets strengthens the credibility of generative models in producing biologically meaningful outcomes.

This study underscores the potential of interpretable, data-driven tools in advancing food safety and quality control. Future studies should focus on applying this framework to other microbial environments and food matrices, thereby extending its utility and validating its generalizability in diverse real-world settings. Future research should also therefore aim to incorporate multi-seasonal and geographically diverse datasets to enhance the robustness and applicability of the model. Moreover, integrating functional omics data—such as metagenomics or metatranscriptomics—could enable

direct assessment of antimicrobial production potential, including the expression of biosynthetic gene clusters responsible for nisin synthesis.

Declarations

- **Funding** This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
- **Conflict of interest** The authors declare no conflict of interest or competing interests related to this work.
- **Ethical confirmation** This study does not involve human participants, animals, or sensitive data that would require ethics approval or consent to participate.
- **Consent for Publication** All authors have read and approved the final manuscript and consent to its publication.
- **Data Availability** The datasets generated and analysed during the current study are available from the corresponding author on reasonable request.
- **Materials Availability** All materials used in this study are provided in detail in the manuscript.
- **Code Availability** The code used in this study is available upon request from the corresponding author.
- **Author Contribution**
  Volkan Ince: Conceptualization, Methodology, Data Processing, Software, Writing—original draft preparation.
  Mohamed Bader-El-Den: Methodology, Writing—review and editing, Supervision.
  Omer Faruk Sari: Conceptualization, Methodology, Writing—review and editing.
  Annette Sansom: Data Collection,Industrial Perspective, Practical Insights, Feedback.
  All authors have read and agreed to the published version of the manuscript.

Acknowledgments

The authors gratefully acknowledge Campden BRI for supplying the dataset utilized in this study. We also sincerely thank the University of Portsmouth for its support and the resources that made this research possible. Additionally, we extend our appreciation to our colleagues and collaborators for their insightful feedback and constructive contributions throughout the course of this work.

References

Arora, A., Arora, A.: Generative adversarial networks and synthetic patient data: current challenges and future perspectives. Future healthcare journal **9**(2), 190–193 (2022)

Al, S., Ciloglu, F.U., Akcay, A., Koluman, A.: Machine learning models for prediction of escherichia coli o157: H7 growth in raw ground beef at different storage temperatures. Meat Science **210**, 109421 (2024)

Breiman, L.: Random forests. Machine learning **45**, 5–32 (2001)

Bi, X., Zhou, Z., Wang, X., Jiang, X., Chen, L., Xing, Y., Che, Z.: Changes in the microbial content and quality attributes of carrot juice treated by a combination of ultrasound and nisin during storage. Food and Bioprocess Technology **13**, 1556–1565 (2020)

BRI: Science and technology for the food and drink industry — campdenbri.co.uk. https://www.campdenbri.co.uk/. [Accessed 30-06-2024]

Casco, M., Jagus, R.J., Agüero, M.V., Fernandez, M.V.: Ultrasound and its combination with natural antimicrobials: Effects on shelf life and quality stability of a fruit and vegetable smoothie. Food and Bioprocess Technology, 1–16 (2022)

Chaves López, C., Serio, A., Montalvo, C., Ramirez, C., Peréz Álvarez, J.A., Paparella, A., Mastrocola, D., Martuscelli, M.: Effect of nisin on biogenic amines and shelf life of vacuum packaged rainbow trout (oncorhynchus mykiss) fillets. Journal of Food Science and Technology **54**, 3268–3277 (2017)

Di Bucchianico, A.: Coefficient of determination (r 2). Encyclopedia of statistics in quality and reliability (2008)

Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189–1232 (2001)

Hiura, S., Koseki, S., Koyama, K.: Prediction of population behavior of listeria monocytogenes in food using machine learning and a microbial growth and survival database. Scientific reports **11**(1), 10613 (2021)

Hodson, T.O.: Root mean square error (rmse) or mean absolute error (mae): When to use them or not. Geoscientific Model Development Discussions **2022**, 1–10 (2022)

Hosoe, J., Sunagawa, J., Nakaoka, S., Koseki, S., Koyama, K.: Data mining for prediction and interpretation of bacterial population behavior in food. Frontiers in Food Science and Technology **2**, 979028 (2022)

Ince, V., Bader-El-Den, M., Alderton, J., Arabikhan, F., Sari, O.F., Sansom, A.: Machine learning-based prediction of clostridium growth in pork meat using explainable artificial intelligence. Journal of Food Science and Technology, 1–14 (2025)

Ince, V., Bader-El-Den, M., Sari, O.F.: Enhanced dataset synthesis using ctgan for metagenomic dataset. In: 2024 IEEE 12th International Conference on Intelligent Systems (IS), pp. 1–6 (2024). IEEE

Khatiman, M.N.A., Abu-Samah, A., Azman, M.A., Nordin, R., Abdullah, N.F.: Generation of synthetic 5g network dataset using generative adversarial network (gan). In: 2023 IEEE 16th Malaysia International Conference on Communication (MICC), pp. 141–145 (2023). IEEE

Kuda, T., Fujii, A., Nakamura, A., Takahashi, H.: Inhibitory effect of lactococcus lactis nagasaki-su6 isolated from algal beach cast on listeria monocytogenes during and after fermentation of cow-and soymilk. Food Bioscience, 103762 (2024)

Kiran, A., Kumar, S.S.: A methodology and an empirical analysis to determine the most suitable

synthetic data generator. IEEE Access **12**, 12209–12228 (2024)

Koyama, K., Kubo, K., Hiura, S., Koseki, S.: Is skipping the definition of primary and secondary models possible? prediction of escherichia coli o157 growth by machine learning. Journal of Microbiological Methods **192**, 106366 (2022)

Kim, H.-J., Kim, H.-J., Jo, C.: A non-destructive predictive model for estimating the freshness/spoilage of packaged chicken meat using changes in drip metabolites. International Journal of Food Microbiology **419**, 110738 (2024)

Kuroda, S., Okuda, H., Ishida, W., Koseki, S.: Modeling growth limits of bacillus spp. spores by using deep-learning algorithm. Food microbiology **78**, 38–45 (2019)

Karlsen, S.T., Vesth, T.C., Oregaard, G., Poulsen, V.K., Lund, O., Henderson, G., Bælum, J.: Machine learning predicts and provides insights into milk acidification rates of lactococcus lactis. PLoS One **16**(3), 0246287 (2021)

Lu, Y., Chen, D., Olaniyi, E., Huang, Y.: Generative adversarial networks (gans) for image augmentation in agriculture: A systematic review. Computers and Electronics in Agriculture **200**, 107208 (2022)

Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)

Li, M., Li, Y., Huang, X., Zhao, G., Tian, W.: Evaluating growth models of pseudomonas spp. in seasoned prepared chicken stored at different temperatures by the principal component analysis (pca). Food microbiology **40**, 41–47 (2014)

Mert, A.: Enhanced dataset synthesis using conditional generative adversarial networks. Biomedical Engineering Letters **13**(1), 41–48 (2023)

Munck, N., Njage, P.M.K., Leekitcharoenphon, P., Litrup, E., Hald, T.: Application of whole-genome sequences and machine learning in source attribution of salmonella typhimurium. Risk Analysis **40**(9), 1693–1705 (2020)

Marcelli, V., Osimani, A., Aquilanti, L.: Research progress on the use of lactic acid bacteria as natural bio-preservatives against pseudomonas spp. in meat and meat products: A review. Food Research International, 115129 (2024)

Mohsina, K., Ratkowsky, D.A., Bowman, J.P., Powell, S., Kaur, M., Tamplin, M.L.: Effect of glucose, ph and lactic acid on carnobacterium maltaromaticum, brochothrix thermosphacta and serratia liquefaciens within a commercial heat-shrunk vacuum-package film. Food microbiology **91**, 103515 (2020)

Merchán, A.V., Ruiz-Moyano, S., Benito, M.J., Hernández, M.V., Cabañas, C.M., Román, Á.C.: Metabarcoding analysis reveals a differential bacterial community profile associated with 'torta del casar'and 'queso de la serena'pdo cheeses. Food Bioscience **57**, 103491 (2024)

Nti, I.K., Nyarko-Boateng, O., Aning, J., *et al.*: Performance of machine learning algorithms with different k values in k-fold cross-validation. International Journal of Information Technology and Computer Science **13**(6), 61–71 (2021)

Nath, K.G., Pandiselvam, R., Sunil, C.: High-pressure processing: Effect on textural properties of food-a review. Journal of Food Engineering **351**, 111521 (2023)

Oliveira, R.B., Ramos, G.L., Sá, P.B., Pereira, A.P.M., Conceição, D.A., Cruz, A.G., Alvarenga, V.O., Sant'Ana, A.S.: Controlling clostridium sporogenes spoilage of "requeijão cremoso" processed cheese: Modeling the growth/no-growth probability as a function of ph, sodium chloride and nisin. Food Control **162**, 110435 (2024)

Phuapaiboon, P., Leenanon, B., Levin, R.E.: Effect of lactococcus lactis immobilized within pineapple and yam bean segments, and jerusalem artichoke powder on its viability and quality of yogurt. Food and Bioprocess Technology **6**, 2751–2762 (2013)

Piredda, R., Mottola, A., Cipriano, G., Carlucci, R., Ciccarese, G., Di Pinto, A.: Next generation sequencing (ngs) approach applied to species identification in mixed processed seafood products. Food Control **133**, 108590 (2022)

Papoutsoglou, G., Tarazona, S., Lopes, M.B., Klammsteiner, T., Ibrahimi, E., Eckenberger, J., Novielli, P., Tonda, A., Simeon, A., Shigdel, R., *et al.*: Machine learning approaches in microbiome research: challenges and best practices. Frontiers in microbiology **14**, 1261889 (2023)

Qu, T., Wang, P., Zhao, X., Liang, L., Ji, Q., Ge, Y., Chen, Y.: Metagenomic profiles of the antimicrobial resistance in traditional chinese fermented meat products: Core resistome and co-occurrence patterns. International Journal of Food Microbiology **418**, 110740 (2024)

SDV: TVAESynthesizer — Synthetic Data Vault — docs.sdv.dev. https://docs.sdv.dev/sdv/single-table-data/modeling/synthesizers/tvaesynthesizer. [Accessed 25-03-2025]

Song, Y., Liang, J., Lu, J., Zhao, X.: An efficient instance selection algorithm for k nearest neighbor regression. Neurocomputing **251**, 26–34 (2017)

Spanu, C., Piras, F., Mocci, A., Nieddu, G., De Santis, E., Scarano, C.: Use of carnobacterium spp protective culture in map packed ricotta fresca cheese to control pseudomonas spp. Food Microbiology **74**, 50–56 (2018)

Sun, R., Vermeulen, A., Devlieghere, F.: Modeling the combined effect of temperature, ph, acetic and lactic acid concentrations on the growth/no growth interface of acid-tolerant bacillus spores. International Journal of Food Microbiology **360**, 109419 (2021)

Tarlak, F.: Machine learning-based software for predicting pseudomonas spp. growth dynamics in culture media. Life **14**(11), 1490 (2024)

Topçuoğlu, B.D., Lesniak, N.A., Ruffin IV, M.T., Wiens, J., Schloss, P.D.: A framework for effective application of machine learning to microbiome-based classification problems. MBio **11**(3), 10–1128 (2020)

Toomik, E., Rood, L., Bowman, J.P., Kocharunchitt, C.: Microbial spoilage mechanisms of vacuum-packed lamb meat: A review. International Journal of Food Microbiology **387**, 110056 (2023)

Tarlak, F., Yücel, Ö.: Prediction of pseudomonas spp. population in food products and culture media using machine learning-based regression methods. Life **13**(7), 1430 (2023)

Wang, Z., Bovik, A.C.: Mean squared error: Love it or leave it? a new look at signal fidelity measures. IEEE signal processing magazine **26**(1), 98–117 (2009)

Wang, M., Kong, Q., Yu, H., Jiang, M., Han, Z.: Identification of dominant spoilage bacteria, bacterial diversity analysis, and antibacterial exploration of chilled monkfish. Food Bioscience **56**, 103121 (2023)

Watanabe, S., Masamura, N., Satoh, S.-y., Hirao, T.: Technique for the identification of insect species in processed foods based on three short dna sequences. Food Control **153**, 109908 (2023)

Yap, M., O'Sullivan, O., O'Toole, P.W., Sheehan, J.J., Fenelon, M.A., Cotter, P.D.: Seasonal and geographical impact on the irish raw milk microbiota correlates with chemical composition and climatic variables. MSystems **9**(4), 01290–23 (2024)

Yücel, Ö., Tarlak, F.: An intelligent based prediction of microbial behaviour in beef. Food Control **148**, 109665 (2023)

Yildirim-Yalcin, M., Yucel, O., Tarlak, F.: Development of prediction software to describe total

mesophilic bacteria in spinach using a machine learning-based regression approach. Food Science and Technology International **31**(1), 3–10 (2025)

Zain, A.M., Haron, H., Qasem, S.N., Sharif, S.: Regression and ann models for estimating minimum value of machining performance. Applied Mathematical Modelling **36**(4), 1477–1492 (2012)

Zargari Marandi, R.: Explainer: an r package to explain machine learning models. Bioinformatics advances **4**(1), 049 (2024)

Zhao, X., Shi, C., Meng, R., Liu, Z., Huang, Y., Zhao, Z., Guo, N.: Effect of nisin and perilla oil combination against listeria monocytogenes and staphylococcus aureus in milk. Journal of food science and technology **53**, 2644–2653 (2016)

Zhang, D., Zhang, L., Cong, B., Hu, J., Zhang, X.: Multi-frequency bioimpedance combined with machine learning for frozen-thawed quality evaluation of atlantic salmon. Journal of Food Engineering **382**, 112195 (2024)